Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 6297–6301, June 1996
Genetics

A bacterial artificial chromosome-based framework contig map of human chromosome 22q

UNG-JIN KIM*†, HIROAKI SHIZUYA*, HYUNG-LYUN KANG*‡, SUN-SHIM CHOI*§, CHARMAIN L. GARRETT¶, LUC J. SMINK¶, BRUCE W. BIRREN*‖, JULIE R. KORENBERG**, IAN DUNHAM¶, AND MELVIN I. SIMON*†

*Division of Biology, 147–75, California Institute of Technology, Pasadena, CA 91125; ¶The Sanger Center, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, United Kingdom; and **Medical Genetics Birth Defects Center, Ahmanson Department of Pediatrics, Cedars–Sinai Medical Center, University of California, Los Angeles, CA 90048

Contributed by Melvin I. Simon, February 26, 1996

ABSTRACT **We have constructed a physical map of human chromosome 22q using bacterial artificial chromosome (BAC) clones. The map consists of 613 chromosome 22-specific BAC clones that have been localized and assembled into contigs using 452 landmarks, 346 of which were previously ordered and mapped to specific regions of the q arm of the chromosome by means of chromosome 22-specific yeast artificial chromosome clones. The BAC-based map provides immediate access to clones that are stable and convenient for direct genome analysis. The approach to rapidly developing marker-specific BAC contigs is relatively straightforward and can be extended to generate scaffold BAC contig maps of the rest of the chromosomes. These contigs will provide substrates for sequencing the entire human genome. We discuss how to efficiently close contig gaps using the end sequences of BAC clone inserts.**

To date, all of the human chromosomes have been mapped on the basis of a variety of markers and resources including sequence-tagged sites (STSs) and yeast artificial chromosomes (YACs; refs. 1–8). Currently, large numbers of STSs and expressed sequence tags (ESTs) are being generated and localized on human chromosomes, and it is expected that a human genome map with markers spaced at 100-kb intervals will soon be available (9). YACs with inserts >1000 kb have provided an efficient means to develop genome-wide contig maps that integrate a variety of markers. However, YAC-based physical maps are not ideal for direct genome sequencing, because YAC clones are often plagued with chimerism, rearrangement, and deletion (10–12) and because isolating reasonable amounts of pure YAC DNA for molecular analysis is difficult. Cosmid contigs have been used to obtain high-resolution maps. However, because cosmids carry relatively short inserts, map assembly and large-scale sequencing are not very efficient. We describe an approach to facilitate large-scale genomic sequencing, which involves transforming YAC-based maps of human chromosomes into maps based on large insert bacterial clones, such as bacterial artificial chromosomes (BACs), which are stable, represent the largest human inserts obtainable amongst bacterial clones, and can readily be manipulated and directly used as sequencing substrates.

The BAC system employs a vector based on the *Escherichia coli* F-factor replicon that maintains clones at single-copy in recombination-deficient bacterial hosts (13). We have constructed extensive genomic BAC libraries both for the human (unpublished data) and the mouse genomes (unpublished data), with inserts as large as 350 kb. Further, we have demonstrated the stability of large human DNA inserts in BAC vectors during extended propagation of the clones (13, 14). Because of their stability, relatively large size, simple purification, general ease of screening using hybridization or PCR to correlate markers with corresponding clones, BACs provide reliable and efficient materials for the construction of sequence-ready maps. We report here an approach to rapidly constructing a chromosome-scale BAC scaffold map by screening a human BAC library with chromosome 22-specific markers. Many of these markers were previously ordered via YAC-based chromosome 22q mapping (7). We discuss how the map can directly be used for the initiation of large-scale genome sequencing. This approach serves as a paradigm for the rapid development of sequence-ready physical maps for other chromosomes and for clones that can be used as substrates to sequence the entire human genome.

MATERIALS AND METHODS

A human BAC library with ≈4-fold coverage was constructed from human fibroblast primary cell line (ATCC CRL 1905) and screened either by colony hybridization or by PCR analysis of library subpools (unpublished data). The 452-chromosome 22-specific markers that were used for screening the library are listed in Table 1. The anonymous STSs, ESTs, YAC end probes, and fluorescence *in situ* hybridization (FISH)-mapped cosmids used in this experiment were described (7). FISH mapping of chromosome 22-specific fosmids (15) and BACs (16) has been described elsewhere. To screen the library with cosmids and fosmids, inserts were isolated from low-melt gels after *Sfi*I or *Not*I digestion, respectively, and radiochemically labeled as described (16, 17). STS primers generated by the genomic group at the Whitehead Institute (Cambridge, MA) have been described elsewhere (8).

The insert size of the BAC clones was determined by pulsed-field gel electrophoresis of *Not*I digests of miniprep BAC DNA (13). Contaminating *E. coli* DNA in the miniprep can efficiently be removed by using ATP-dependent DNAseI (Epicentre Technologies, Madison, WI) that selectively digests linear DNA and leaves covalently closed circular DNA intact (data not shown). The map was assembled by assigning positive BACs identified by hybridization or PCR screening to corresponding markers according to the order shown in the YAC map (7) and by placing other chromosome 22-specific markers at appropriate positions on BAC clones or contigs. In drawing the map, markers were ordered on the basis of assignments to BAC clones within a contig. The contiguity of BACs was

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: STS, sequence-tagged site; YAC, yeast artificial chromosome; EST, expressed sequence tag; BAC, bacterial artificial chromosome; FISH, fluorescence *in situ* hybridization.

†To whom reprint requests should be addressed.
‡Permanent address: Department of Microbiology, Seoul National University, Seoul 151–742, Korea.
§Permanent address: Department of Life Science, Pohang Institute of Science and Technology, Pohang 790–784, Korea.
‖Present address: Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, 1 Kendall Square, Building 300, Cambridge, MA 02139.

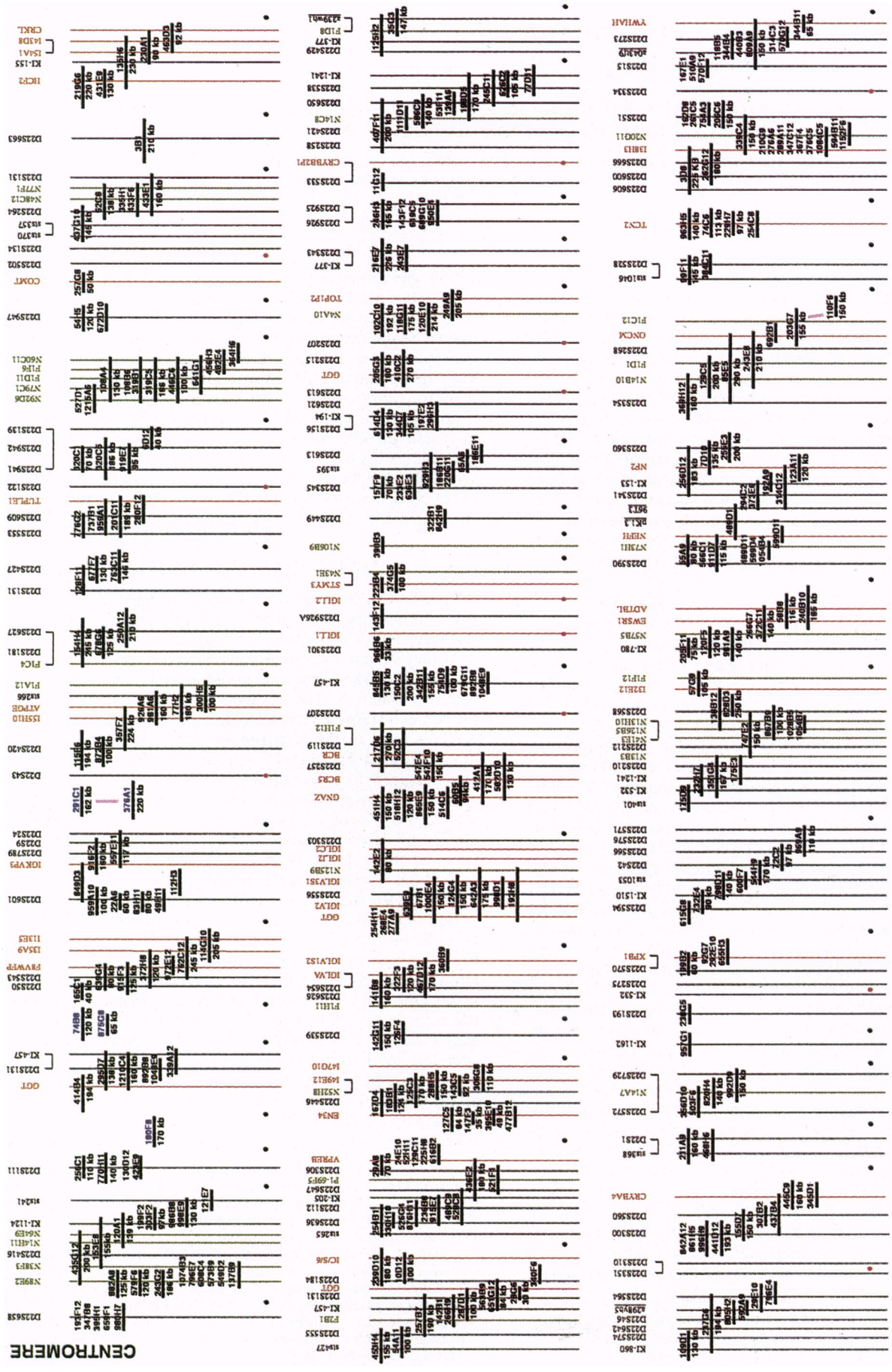
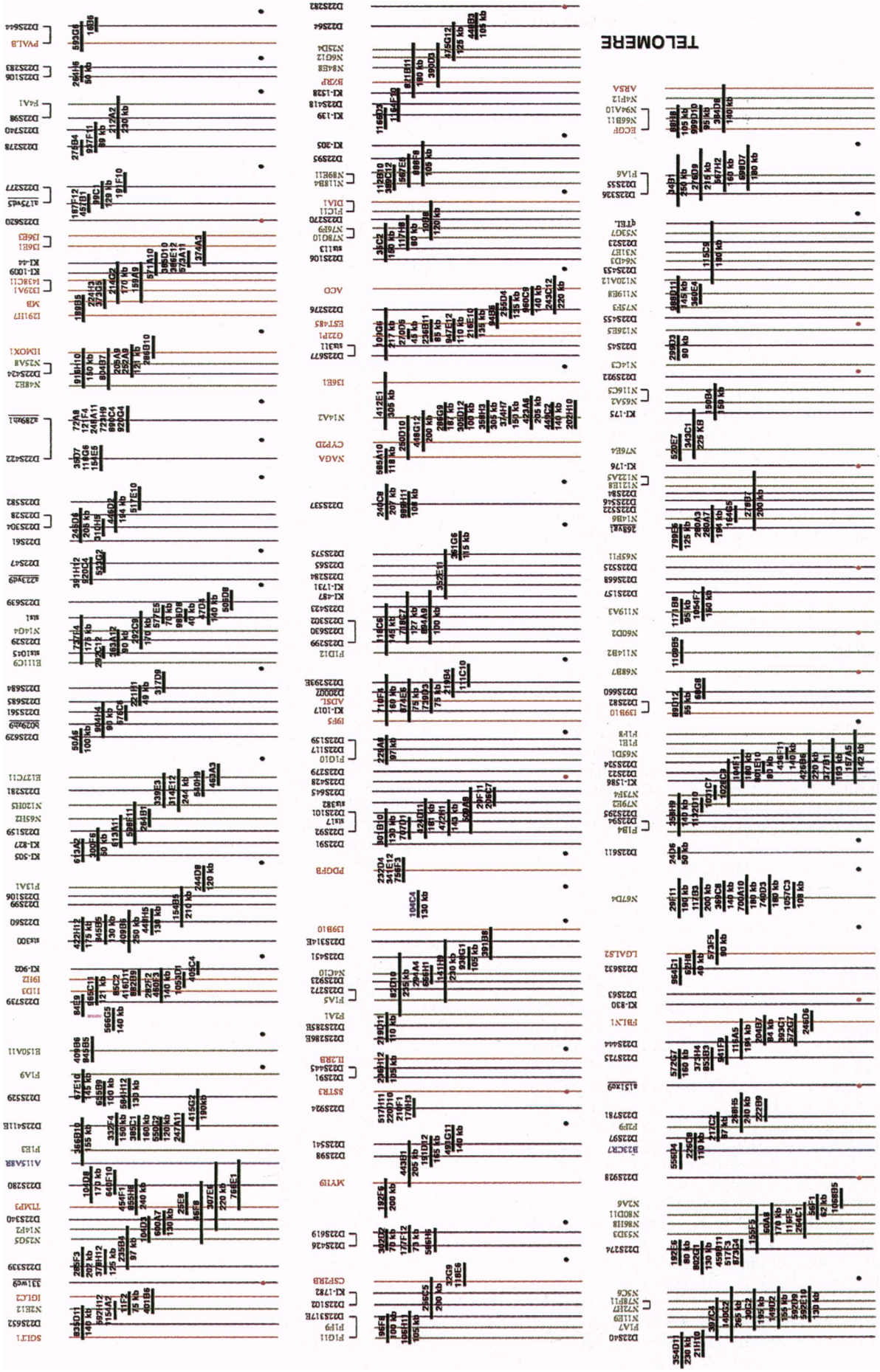


FIG. 1. A framework BAC contig map of human chromosome 22q. Shown at the top of the map are total 452 markers, which have been used to screen the 4-fold BAC library. The markers were as follows: D22S prefix, anonymous markers; KI prefix, anonymous markers from Karolinska Institute (21); sts prefix, anonymous STS from Whitehead Institute (22); E and N prefix cosmid and F prefix fosmid, clones localized to chromosome 22 subregions by FISH mapping (7, 16); I prefix, chromosome 22-specific cDNA clones from the normalized human cDNA libraries of I.M.A.G.E. consortium (23); A and B prefix, the YAC-end probes (7); and P1-prefix P1 clones. Underlined markers indicate microsatellites (CA-repeats). Color codes indicate as follows: black, anonymous markers; deep red, markers derived from genes or pseudogenes; and green, cosmids and fosmids. The order of the markers, including most of the anonymous markers, YAC ends, and cosmids, was determined primarily by the previous YAC-based chromosome 22q map (7). Horizontal brackets indicate the pair of markers that could not be ordered with respect to each other (Fig. 1 *legend continues on opposite page.*)



either on the basis of the current BA contig map or previous YAC map. A total of 613 BAC clones are shown on thick horizontal bars with insert sizes, where they are known, indicated underneath. Multiple BAC clones piled on a single horizontal bar indicate that they are positive for the same set of markers, and their order and insert sizes have not been determined. BAC clones that have been localized to the specific positions previously (7) are indicated in blue. Clones that hybridize positively to each other are shown connected to each other by thick pale purple lines. The relative positions of the markers with respect to BAC contigs are indicated by vertical lines that extend from the markers and cross BAC clones. Confirmed and putative gaps are indicated by 26 red dots and 84 black dots along the near bottom of the map, respectively. Objects including the spaces between markers and contigs, and the size of BACs do not reflect actual scale.

Table 1. Summary of the markers shown in the chromosome 22q BAC scaffold map

Markers	Landmarks, no.
Anonymous markers	
D22S	203
KI	30
sts	18
Others	14
Total	265
ESTs	57
cDNAs	21
YAC ends	2
FISH-mapped cosmids	80
FISH-mapped fosmids	26
P1	1
Total	452

All anonymous markers except KI probes and all ESTs were primer pairs. The rest of the markers were hybridization probes.

confirmed separately (data not shown) by the BAC fingerprinting procedure (18). The map shows the order and contiguity of BAC clones and markers with respect to each other, and the position of the physical gaps between contigs. However, the distance between the objects such as markers, clones, and contigs and the extension of the BAC clones do not reflect actual scale. Information regarding BAC libraries and chromosome 22-specific clones is available from our web page (19). All of the BAC clones identified in this paper are available from Research Genetics (Huntsville, AL).

RESULTS AND DISCUSSION

Table 1 summarizes a total of 452 markers that were used. These markers were positioned on the map according to the results of the screening of a 4-fold library (Fig. 1). Three hundred forty-six of the markers on the map were derived from the YAC map and thus served as anchor points for the current map. Of the >800 putative chromosome 22-specific BACs selected by various methods, 613 clones could be placed on the map unequivocally. The rest of the BACs that were not included in the map are either positive for known chromosome 22-specific landmarks that have not been precisely mapped or the BACs mapped to chromosome 22q only by means of the FISH-mapping technique. Twenty BACs showed positive signals with the GGT marker, which is repeated five times in the 22q11 region. Six of these clones did not correspond to any other distinctive landmarks that have been mapped and were not placed on the map, except for the contig with 414B4, which was FISH-mapped to the centromeric region of chromosome 22 and was placed in the most centromeric GGT repeat. A pair of GGT repeats that are very close to each other in the YAC map could not be resolved. The contiguity of the clones within of each of the contigs was determined by restriction fingerprinting (see below).

Only 35 of 452 markers tested produced no corresponding BAC; thus 92.3% of the loci were covered by at least one BAC. This is in good agreement with ≈92% genomic coverage expected of the 4-fold BAC library (unpublished data). Because the BAC contigs cover nearly 90% of the YAC map (data not shown), which itself represents ≈93% of the chromosome 22q arm (7), we estimate that the actual coverage of the BAC map should be ≈80% of the q-arm of the chromosome. This estimate is also in good agreement with the physical measurements from the size and the number of BAC clones on the map. Overall redundancy of BAC contigs is ≈2.2-fold, as determined by averaging the number of markers assigned to each of the BACs. The average size of 309 BACs shown in the map with known insert sizes, which represent random samples of the 613 mapped BAC clones, is 145 kb. Therefore, it is possible to extrapolate that the total length of the mapped BACs would be ≈613 times 145 kb, yielding 88,885 kb. The estimated coverage of the chromosome 22q can be inferred by dividing the total length by the redundancy factor 2.2, which yields 40.4 Mb, ≈80% of the estimated 50 Mb chromosome 22q arm. Therefore, the average size of 111 BAC contigs should be ≈364 kb, with an average gap size of 91 kb that may vary considerably depending on the actual size of the chromosome 22q arm. On the average, each of the contigs consists of 5.5 BACs.

Contigs initially established by the content of markers on BAC clones were further characterized. Members of a contig identified by a set of markers were subjected to restriction fingerprint analysis for the confirmation of the contiguity among the clones using the method originally developed by Sulston *et al.* (20, 24), which has been modified for BAC fingerprinting (18). Overall, the order of markers established by the YAC-based map was well conserved with some exceptions. In the process of contig assembly, a number of changes (18 cases) were introduced in the local order of the markers (data not shown). There were 12 cases where the order of two or three consecutive markers that could not be ordered by YACs could be determined by BACs (data not shown). In addition, four markers (KI-205, KI-487, KI-505, and KI-827) were positive to BAC contigs that localized to different loci than previously determined in the YAC map.

Gaps between neighboring contigs were considered confirmed if they were separated by one or more markers that did not identify a corresponding BAC. The other apparent gaps may represent our inability to detect contiguity, and these were considered as putative gaps. In the current map, there are 26 confirmed gaps and 84 putative gaps. We estimate that most of the gaps will be smaller than the average BAC clone, thus many of them could be closed by single BACs. Gaps can be filled by generating new probes that are specific for contig ends. First, the DNA sequences at the ends of the BAC inserts at the boundaries of the contigs can be determined by directly using the BAC miniprep DNA as sequencing templates (unpublished data). Then the sequences can be used to design new markers to screen a different large insert library and to obtain clones that will extend the contigs and close the gaps. A library with 15-fold coverage of the human genome in large insert bacterial clones would allow us to assemble a BAC map with almost complete coverage. However, for most applications, complete coverage may not be necessary. In its current state, the map can be used for positional cloning and targeted genomic sequencing. One can readily choose a set of BAC clones from the map that can be used for sequencing a specific genomic locus. In fact, a number of BACs shown in our map are being sequenced by the Sanger Center (I.D., unpublished data).

To efficiently sequence the entire chromosome 22q arm, one could select a set of ≈200–250 nonoverlapping or minimally overlapping BACs directly from the map, which would then be sequenced. To close the gaps between the sequenced islands, both ends of the clone inserts from all known chromosome 22-specific BACs would then be sequenced using BAC miniprep DNA directly as sequencing templates. With BAC end sequence information, one can precisely determine the extent of the overlaps of these clones with the sequence-completed BACs. A suggestion has been made that the entire human genome could be sequenced distributively by first generating end sequences of large numbers of random BAC clones—e.g., 200,000 BAC clones corresponding to ≈10-fold genomic coverage—given that the average insert size is ≈150 kb (C. Venter, personal communication). The BAC end sequences would then be used to align these BACs with sequence-complete BACs to select minimally overlapping clones for further sequencing.

By screening a moderate-depth BAC library using various markers from the previously developed YAC map, we could efficiently construct a highly representative physical contig map for the q arm of human chromosome 22. On average, a contig covering >200 kb is obtained each time a 4-fold library is screened with an EST or an STS marker. Thus by applying the approach that we have described to the rest of the human chromosomes, where mapped EST or STS markers at better than 200-kb resolution already exist, it will be possible to rapidly build scaffold BAC contig maps for those chromosomes. The scaffold could be further refined and extended for more complete coverage of the genome using the BAC end-sequencing strategy. It can also be used as a guide to provide substrates for efficient large-scale sequencing of the entire human and murine genomes.

We thank Simon Foote, Thomas Hudson, and Eric Lander for providing us with dozens of chromosome 22-specific STS primers, Charles Auffrey and Greg Lennon for chromosome 22-specific cDNAs, Jan Dumanski for KI probes, John Collins for comments on the manuscript, Ulrich Weier for the P1 probe, and Tatiana Slepak, April Mengos, and Daniel Eckstein for excellent technical assistance. This work was supported by the U.S. Department of Energy Grant FG0389ER60891.

1. Foote, S., Vollrath, D., Hilton, A. & Page, D. C. (1992) *Science* **258,** 60–66.
2. Chumakov, I., Rigault, P., Guillous, S., Ougen, P., Billaut, A., *et al.* (1992) *Nature (London)* **359,** 380–387.
3. Chumakov, I., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billaut, A., *et al.* (1995) *Nature (London)* **377** (Suppl.), 177–297.
4. Gemmill, R. M., Chumakov, I. M., Scott, P., Waggoner, B., Rigault, P., *et al.* (1995) *Nature (London)* **377** (Suppl.), 299–319.
5. Krauter, K., Montgomery, K., Yoon, S.-J., LeBlanc-Straceski, J., Renault, B., *et al.* (1995) *Nature (London)* **377** (Suppl.), 321–333.
6. Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., *et al.* (1995) *Nature (London)* **377** (Suppl.), 335–365.
7. Collins, J. E., Cole, C. G., Smink, L. J., Garrett, C. L., Leversha, M. A., *et al.* (1995) *Nature (London)* **377** (Suppl.), 367–379.
8. Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J. L., Castle, A. B., *et al.* (1995) *Science* **270,** 1945–1954.
9. Stewart, A. (1995) *Genome Digest* **2,** 6–9.
10. Kouprina, N., Eldarov, M., Moyzis, R. & Larionov, V. (1994) *Genomics* **21,** 7–17.
11. Larionov, V., Kouprina, N., Nikolaishvili, N. & Resnick, M. A. (1994) *Nucleic Acids Res.* **22,** 4154–4162.
12. Green, E. D., Riethman, H. C., Dutchik, J. E. & Olson, M. V. (1991) *Genomics* **11,** 658–669.
13. Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y. & Simon, M. I. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8794–8797.
14. Kim, U.-J., Shizuya, H., de Jong, P., Birren, B. & Simon, M. I. (1992) *Nucleic Acids Res.* **20,** 1083–1085.
15. Birren, B. W., Tachi-iri, Y., Kim, U-J., Nguyen, M., Shizuya, H., Korenberg, J. R. & Simon, M. I. (1996) *Genomics*, in press.
16. Kim, U.-J., Shizuya, H., Chen, X.-N., Deaven, L., Speicher, S., Solomon, J., Korenberg, J. & Simon, M. I. (1995) *Genet. Anal. Biomol. Eng.* **12,** 73–79.
17. Kim, U.-J., Shizuya, H., Birren, B., Slepak, T., de Jong, P. & Simon, M. I. (1994) *Genomics* **22,** 336–339.
18. Kim, U.-J., Amemiya, C., Evan, G. A. & Birren, B. W. (1996) *Bacterial Cloning Vectors in Genome Analysis: Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), in press.
19. Kim, U.-J., Shizuya, H., Kang, H.-L., Choi, S.-S., Garrett, C. L., Smink, L. J., Birren, B. W., Korenberg, J. R., Dunham, I. & Simon, M. I., http://www.tree.caltech.edu.
20. Sulston, J. E., Mallet, F., Durbin, R. & Horsnell, T. (1988) *Comput. Appl. Biosci.* **5,** 101–106.
21. Dumanski, J. P., Geurts van Kessel, Ad. H. M., Ruttledge, M., Wladis, A., Sugawa, N., Collins, V. P. & Nordenskjöld, M. (1991) *Hum. Genet.* **84,** 219–222.
22. Bell, C. J., Budarf, M. L., Nieuwenhuijsen, B. W., Barnoski, B. L., Buetow, K. H., *et al.* (1995) *Hum. Mol. Genet.* **4,** 59–69.
23. Auffrey, C., Behar, G., Bois, F., Bouchier, C., Da Silva, C., Devignes, M.-D., Duprat, S., Houlgatte, R., Jumeau, M.-N., Lamy, B., Lorenzo, F., Mitchell, H., Mariage-Samson, R., Pietu, G., Pouliot, Y., Sebastiani-Kabaktchis, C. & Tessier, A. (1995) *C. R. Acad. Sci. Ser. 3* **318,** 263–272.
24. Sulston, J. E., Mallet, F., Staden, R., Durbin, R., Horsnell, T. & Coulson, A. (1988) *Comput. Appl. Biosci.* **4,** 125–132.